Protecting Privacy of Sensitive Value Distributions in Data Release

M. Bezzi[1], S. De Capitani di Vimercati[2], G. Livraga[2], P. Samarati[2]

[1] SAP Research, Sophia-Antipolis, France
michele.bezzi@sap.com
[2] Università degli Studi di Milano, 26013 Crema, Italy
{sabrina.decapitani, pierangela.samarati}@unimi.it,
giovanni.livraga@guest.unimi.it

Abstract. In today's electronic society, data sharing and dissemination are more and more increasing, leading to concerns about the proper protection of privacy. In this paper, we address a novel privacy problem that arises when non sensitive information is incrementally released and sensitive information can be inferred exploiting dependencies of sensitive information on the released data. We propose a model capturing this inference problem where sensitive information is characterized by peculiar distributions of non sensitive released data. We also discuss possible approaches for run time enforcement of safe releases.

1 Introduction

Sharing and dissemination of information play a central role in today's information society. Governmental, public, and private institutions are increasingly required to make their data electronically available, as well as to offer services and data access over the Internet. This implies disclosing to external parties or sharing information once considered classified or accessible only internally, that must now be made partially available to outside interests. Such information release, publication and dissemination are clearly selective. Data maintained by any organization may in fact considerably differ with respect to the needs for sharing with external parties as well as for their sensitivity. Data publication and sharing must then ensure on one hand the satisfaction of possible needs for data to be fed to external parties and on the other hand, proper protection of sensitive data to preserve the confidentiality and/or the privacy of involved individuals. The problem is notably complex, since the possible correlations and dependencies existing among data can introduce inference channels causing leakage of sensitive information even if such information is not explicitly released. The problem has been under the attention of researchers for decades and a large body of research has addressed different facets of the problem with different settings and assumptions. Such a large body of research includes: statistical databases and statistical data publications (e.g., [1]); multilevel database systems with the problem of establishing proper classification of data, capturing

data relationship and corresponding inference channels (e.g., [6,15]); novel privacy problems introduced by the release of data referring to individuals whose identities or whose associated sensitive information should be maintained private (e.g., [4,5]); protection of associations among data due to possible mining (e.g., [2]); protection of special type of sensitive information (e.g., [3,10,11]). Different approaches have then been proposed addressing all these aspects of the complex privacy problem and offering solutions to block or limit the exposure of possible sensitive or private information. Still new data publication scenarios together with richness of published data and available data sources raise novel problems that need to be addressed.

In this paper, we address a specific problem related to inferences arising from the dependency of sensitive (not released) information referred to some entities, which can be enabled by the observation of other properties regarding such entities. In particular, we are concerned with the possible inferences that can be withdrawn by observing the distribution of values of non sensitive information associated with the entities. For instance, the distribution of soldiers' age in a military location can allow inferring the nature of the location itself, whether it is a headquarter (hosting old officials) or a training campus (hosting young privates). Intuitively, such a problem of sensitive information derivation becomes more serious as the amount of released data increases. In fact, as the amount of data released increases, the confidence in the external observations will increase; also, external observations will tend to be more representative of the real situations. Our problem resembles in some aspects the classical, and much complex, problem of controlling horizontal aggregation of data but it differs from it in several assumptions. In particular, we assume a scenario where an external observer could gather the data released to legitimate users and inference is due to peculiar data values distributions. Also, we are not only concerned with protecting sensitive information associated with specific entities, but also avoiding possible false positives, where sensitive values may improperly be associated (by the observers) with specific entities.

The remainder of this paper is organized as follows. First, we characterize a novel scenario of inference in data publication raising from a real case study that needed consideration (Section 2). Second, we provide a model for capturing when inference can occur in such scenario, providing metrics for evaluating information exposure (Sections 3 and 4). Third, we discuss possible approaches to control data disclosure to ensure that releases are safe with respect to inference channels improperly exposing sensitive information (Section 5).

2 Motivation and reference scenario

We consider a scenario (see Figure 1) where a *data holder* maintains a collection of records stored in a trusted environment. Each record contains different attributes and can be released to authorized parties requiring it. While the records individually taken are not sensitive, their aggregation is considered sensitive since it might enable inferring sensitive information not appearing in the records and

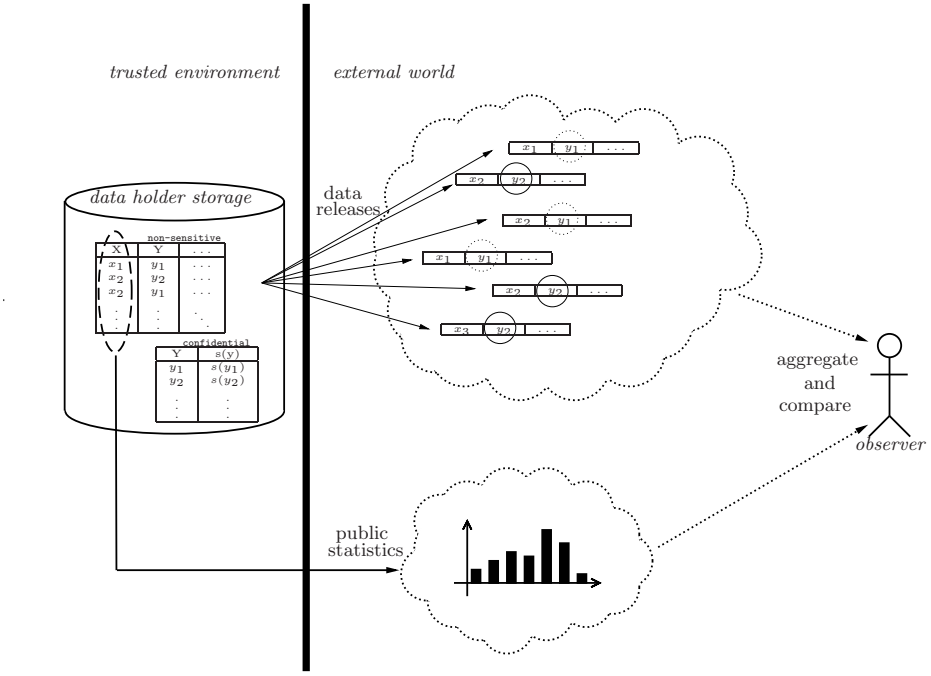


Fig. 1: Reference scenario

not intended for release. We assume all requests for records to be genuine and communication of responses to record release requests to be protected. However, once records are released the data holder has no control on them and therefore *external observers* can potentially gather all the records released. We then consider the worst case assumption of an observer that could be able to retrieve the complete collection of released records (which can happen, for example, if parties to which records are released would make use of a third party provider for storage). We assume that an observer is not aware of the requests submitted to the data holder for retrieving records as well as of the number of records stored at the data holder site.

Our problem is ensuring that the collection of records released to the external world be safe with respect to potential inference of sensitive (not released) information that could be possible by aggregating the released records. We consider a specific case of horizontal aggregation and inference channel due to the distribution of values of certain attributes with respect to other attributes. In particular, inference is caused by a distribution of values that deviates from expected distributions, which are considered as typical and are known to the observers.

In the paper, we refer our examples to a real case scenario characterized as follows. The data holder is a military organization which maintains records on

its personnel. Each record refers to a soldier and reports the attributes `Name`, `Age`, and `Location` where the soldier is on duty. Some of the military locations are headquarters of the army. The information that a location is a headquarter is considered sensitive and neither appears in the soldiers' records nor it is released in other forms. Soldiers' records can be released upon request of the soldiers as well as of external parties (e.g., an external hospital). In addition, the data holder, to be compliant with legal requirements, publicly makes available statistics on the age of the soldiers. The age distribution publicly released, computed on the overall population regardless of the specific locations where soldiers are based, is a distribution that can be considered common and, in general, typically expected at each location. However, locations where headquarters are based show a different age distribution, characterized by an unusual peak of soldiers middle age or older. Such a distribution clearly differs from the expected age distribution, where the majority of soldiers are in their twenties or thirties. The problem is therefore that while single records are considered non sensitive, an observer aggregating all the released records could retrieve the age distribution of the soldiers in the different locations and determine possible deviations from the expected age distribution for certain locations, thus inferring that a given location hosts a headquarter. Our problem consists in ensuring that the release of records to the external world be safe with respect to such inferences.

3 Data model and problem definition

In this section, we provide the notation and formalization of our problem. While our approach is applicable to a generic data model with which the data stored at the data holder site could be organized, for concreteness, we assume data to be maintained as a relational database. The data collection is therefore a table T characterized by a given set A of attributes, and each record is a tuple t in the table. Among the attributes contained in the table, we distinguish a set $Y \subset A$ of attributes corresponding to entities that we call *targets*.

Example 1. With respect to our scenario, table T is defined on the set A={`Name`, `Age`, `Location`} of attributes and Y={`Location`}. In our examples, we assume five different locations L_1, L_2, L_3, L_4, and L_5 are represented in the table.

While the identity (values) of entities Y is non sensitive, such entities are also characterized by *sensitive properties*, denoted $s(Y)$, which are not released. In other words, for each $y \in Y$ the associated sensitive information $s(y)$ does not appear in any released record. However, inference on it can be caused by the distribution of the values of some other attributes $X \subseteq A$ for the specific y. We denote with $P(X)$ the set of *relative frequencies* $p(x)$ of the different x values in the domain of X appearing in table T. Also, we denote with $P(X|y)$ the relative frequency of each value in the domain of X appearing in table T and restricted to the tuples for which Y is equal to y. We call this latter the *y-conditioned distribution* of X in T.

Number of tuples

Age	L1	L2	L3	L4	L5	Total
<18	72	26	38	47	73	256
18-19	151	53	82	140	223	649
20-24	539	147	449	505	736	2376
25-29	452	114	370	418	613	1967
30-34	335	213	234	318	501	1601
35-39	321	238	277	332	538	1706
40-44	128	219	122	162	220	851
45-49	20	205	50	49	76	400
50-54	9	71	28	34	31	173
≥55	2	13	2	2	2	21
Total	2029	1299	1652	2007	3013	10000

(a)

P(Age|Location)

Age	L1	L2	L3	L4	L5	any
<18	3.55	2.00	2.31	2.34	2.42	2.56
18-19	7.44	4.08	4.96	6.98	7.40	6.49
20-24	26.56	11.32	27.18	25.16	24.44	23.76
25-29	22.28	8.78	22.40	20.83	20.35	19.67
30-34	16.51	16.40	14.16	15.84	16.63	16.01
35-39	15.82	18.32	16.77	16.54	17.86	17.06
40-44	6.31	16.86	7.38	8.07	7.30	8.51
45-49	0.99	15.78	3.03	2.44	2.52	4.00
50-54	0.44	5.46	1.69	1.69	1.03	1.73
≥55	0.10	1.00	0.12	0.11	0.05	0.21

(b)

Loc	P(Loc)
L_1	20.29
L_2	12.99
L_3	16.52
L_4	20.07
L_5	30.13

(c)

Fig. 2: Number of tuples in table T by `Age` and `Location` (a), *loc*-conditioned distributions P(`Age`|`Location`) over table T (b), and location frequencies (c)

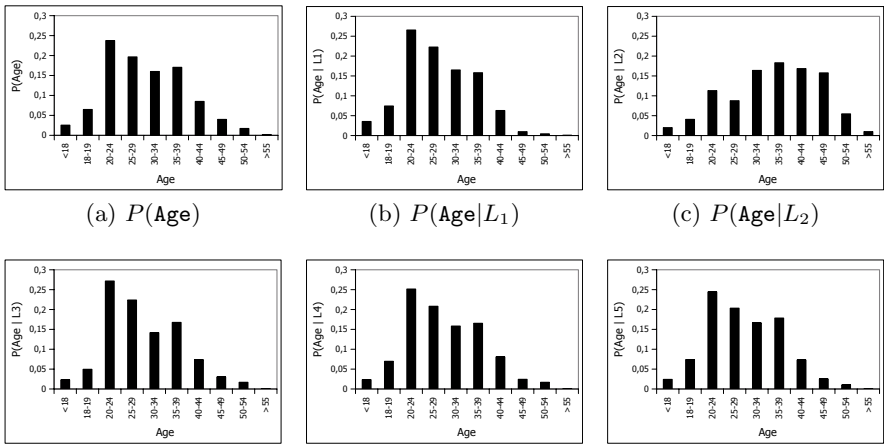
Example 2. In our scenario, $s(Y)$ is the type of the location (e.g., headquarter). The sensitive information $s(y)$ of whether a location y is a headquarter can be inferred from the distribution of the soldier age given the location. Figure 2(a) shows how tuples stored in table T are distributed with respect to the values of attributes `Age` and `Location`. For instance, over the 10000 tuples, 2029 refer to location L_1, 72 of which are of soldiers with age lower than 18. Figure 2(b) reports the corresponding relative frequency of age distributions. In particular, each column *loc*, with $loc \in \{L_1, \ldots, L_5\}$ reports the *loc*-conditioned distribution $P(\texttt{Age}|loc)$ (for convenience expressed in percentage). For instance, it states that 3.55% of the tuples of location L_1 refer to soldiers with age lower than 18. The last column of the table reports the distribution of the age range regardless of the specific location and then corresponds to $P(\texttt{Age})$ (expressed in percentage). Figure 2(c) reports the distribution of soldiers in the different locations regardless of their age (again expressed in percentage). For instance, 20.29% of the 10000 soldiers are based at L_1.

The existence of a correlation between the distribution of values of attributes X for a given target y and the sensitive information $s(y)$ is captured by the definition of *dependency* as follows.

Definition 1 (Dependency). *Let T be a table over attributes A, let X and Y be two disjoint subsets of A, and let $s(Y)$ be a sensitive property of Y. There is a* dependency *between X and Y, denoted $X \rightsquigarrow Y$, if there is a relationship between the conditional distribution $P(X|y)$ and the sensitive information $s(y)$.*

The existence of a dependency between the y-conditioned distribution of X and the sensitive information $s(y)$ introduces an inference channel, since the visibility on $P(X|y)$ potentially enables an observer to infer the sensitive information $s(y)$ even if not released. For instance, with respect to our running example, `Age`$\rightsquigarrow$`Location`.

Definition 1 simply states the existence of a dependency and does not say anything about when a given data distribution causes leakage of the sensitive information. In this paper, we consider the specific case of leakage caused by *peculiar* value distributions that differ from what is considered typical and ex-



(a) $P(\texttt{Age})$ (b) $P(\texttt{Age}|L_1)$ (c) $P(\texttt{Age}|L_2)$

(d) $P(\texttt{Age}|L_3)$ (e) $P(\texttt{Age}|L_4)$ (f) $P(\texttt{Age}|L_5)$

Fig. 3: Baseline distribution (a) and histogram representation of the *loc*-conditioned distributions P(`Age`|`Location`) in Figure 2(b)

pected. We then start to characterize the expected distribution, formally defined as *baseline distribution* as follows.

Definition 2 (Baseline distribution). *Let A be a set of attributes, and let X and Y be two disjoint subsets of A. The* baseline distribution *of X with respect to Y, denoted $B_Y(X)$ is the expected distribution of the different values (or range thereof) of X with respect to Y.*

The baseline distribution is the distribution publicly released by the data holder and can correspond to the real distribution of the values of attributes X in the table T or can be any distribution that the data holder decides to publicly release. We assume the data holder to release truthful data and therefore assume the baseline distribution to coincide with the distribution of the values of X in T, that is, $B_Y(X) = P(X)$. This being said, in the following we simply use $P(X)$ when referring to the baseline distribution.

Example 3. With reference to our example, the baseline distribution $P(\texttt{Age})$ corresponds to the values (expressed in percentage) in the last column of Figure 2(b), which is also graphically reported as a histogram in Figure 3(a). Figures 3(b)-3(f) report the histogram representations of the *loc*-conditioned distributions for the different locations. As clearly visible from the histograms locations, while locations L_1, L_3, L_4, and L_5 enjoy a value distribution that resembles the expected baseline, location L_2 (the headquarter) shows a distribution considerably different.

Inference of sensitive information in our context is therefore caused by unusual distribution of values of X that the observer can learn from viewing released tuples. In the following section, we characterize unusual distributions and

propose an approach to ensure released tuples be protected against such inference. In particular, our goal consists in providing the data holder with a means for assessing whether the release of a tuple (in conjunction with those already released) is safe with to respect to inference or should be denied.[3]

4 Assessing exposure

The first step for determining whether the release of a tuple t concerning a target y (i.e., $y = t[Y]$) is safe consists in characterizing when the y-conditioned distribution of X (i.e., $P(X|y)$) is *peculiar*. In our framing of the problem, this happens when the difference, denoted $\Delta(X, y)$, between $P(X|y)$ and the baseline distribution $P(X)$ characterizes y as an *outlier*, that is, an entity that has a distribution of values of X different from what expected (and from the majority of the other targets). The problem is then how to define such a difference $\Delta(X, y)$. To this purpose, we adopt the classical notion of *Kullback-Leibler distance* D_{KL} and define $\Delta(X, y)$ as follows.

$$\Delta(X, y) = D_{KL}(P(X|y), P(X)) = \sum_{x \in X} p(x|y) \log_2 \frac{p(x|y)}{p(x)} \tag{1}$$

Example 4. Consider the distributions of the `Age` values for the different locations and $P(\texttt{Age})$ in Figure 2(b). We have:

$$\Delta(\texttt{Age}, L_1) = p(< 18|L_1) \log_2 \frac{p(< 18|L_1)}{p(< 18)} + \ldots + p(\geq 55|L_1) \log_2 \frac{p(\geq 55|L_1)}{p(\geq 55)} =$$
$$0.0355 \log_2 \frac{0.0355}{0.0256} + \ldots + 0.0010 \log_2 \frac{0.0010}{0.0021} = 0.12.$$

Similarly, we obtain: $\Delta(\texttt{Age}, L_2) = 0.42$, $\Delta(\texttt{Age}, L_3) = 0.07$, $\Delta(\texttt{Age}, L_4) = 0.06$, and $\Delta(\texttt{Age}, L_5) = 0.06$.

Translating the concept above to the whole table T, we aim at determining the average among the distances of the different y's, each weighted by y's frequency in the table. Such a formula nicely corresponds to the statistical concept of *mutual information*, for which D_{KL} represents a possible decomposition [8]. Intuitively, the mutual information between X and Y characterizes the *average* amount of knowledge about X an observer can have observing Y, or vice versa. The mutual information captures the weighted average of the Kullback-Leibler distance for the different targets as follows.

$$I(X, Y) = \sum_{x \in X, y \in Y} p(y)p(x|y) \ \log_2 \frac{p(x|y)}{p(x)} = \sum_{y \in Y} p(y)\Delta(X, y) \tag{2}$$

[3] Remember that the party requesting the release of the tuple is trusted and the communication is protected. Hence, denying a release does not cause any inference.

Example 5. With respect to our running example, consider the values $p(loc)$, and $\Delta(\texttt{Age}, loc)$, with $loc = L_1, \ldots, L_5$, reported in Figure 2(c) and in Example 4, respectively. We have:
$I(\texttt{Age}, \texttt{Location}) = p(L_1)\Delta(\texttt{Age}, L_1) + p(L_2)\Delta(\texttt{Age}, L_2) + p(L_3)\Delta(\texttt{Age}, L_3) + p(L_4)\Delta(\texttt{Age}, L_4) + p(L_5)\Delta(\texttt{Age}, L_5) = 0.2029 \cdot 0.12 + 0.1299 \cdot 0.42 + 0.1652 \cdot 0.07 + 0.2007 \cdot 0.06 + 0.3013 \cdot 0.06 = 0.12$

The sensitive information $s(y)$ associated with a target $y \in Y$ is considered exposed if $\Delta(X, y)$ deviates from its average $I(X, Y)$ more than a standard deviation σ_Δ. In such a case we say that y is an X-outlier, as defined by the following definition.

Definition 3 (X-outlier). *Let T be a table over attributes A and let X and Y be two subsets of A such that $X \rightsquigarrow Y$. We say that $y \in Y$ is an* X-outlier *if and only if $\Delta(X, y) > I(X, Y) + \sigma_\Delta$, where σ_Δ is the standard deviation of $\Delta(X, y)$.*

Example 6. With respect to our running example, suppose that $\sigma_\Delta = 0.02$ and consider the values of $\Delta(\texttt{Age}, L_1), \ldots, \Delta(\texttt{Age}, L_5)$ in Example 4 and of $I(\texttt{Age}, \texttt{Location})$ in Example 5. L_2 is the unique location that is an Age-outlier since $\Delta(\texttt{Age}, L_2) = 0.42$ is greater than $I(\texttt{Age}, \texttt{Location}) + \sigma_\Delta = 0.14$.

Definition 3 characterizes the actual outliers in the original table T. However, external observers can only see and learn the distribution of values computed on tuples that have been released. By denoting with T_r the set of released tuples and with P_r the value distributions observable on T_r (in contrast to the P observable on T), the knowledge of an external observer can be expressed as the different observations $P_r(X|y)$ she can learn by collecting all the tuples released and the baseline distribution $P(X)$ publicly released by the data holder. We therefore need to characterize the exposure of a target y in terms of how much the observable y-conditioned distribution of X differ from the one expected.

A first term to characterize such exposure is the distance $\Delta_r(X, y)$ of the y-conditioned distribution of X over the released tuples T_r (i.e., $P_r(X|y)$) and the expected baseline distribution (i.e., $P(X)$). A second term that comes into play is the frequency of the specific y in the released dataset T_r. The rational is that since external observers do not have any information about the content of the original table T, they also do not know the number of tuples related to a given y in T; the only information observers can have about a target y is the one observable in T_r. Targets having small frequencies in T_r are then intrinsically more protected than ones having greater frequencies. In fact, if a target y appears with only few occurrences in T_r, an observer is likely not to put great confidence on its distribution, observed over few tuples. For instance, consider a released dataset T_r of 1000 tuples, where 10 tuples refer to y_1 and 990 to y_2, with $P(X|y_1) = P(X|y_2)$. While $\Delta_r(X, y_1)$ will be the same as $\Delta_r(X, y_2)$, an observer might not grant much confidence on the observations on y_1 since they result a limited number of tuples compared to the size of T_r. This aspect is captured by considering the frequency $p_r(y)$ of y in T_r as a weight for the Kullback-Leibler distance when computing the *exposure* of y. We therefore evaluate the exposure for a target y given a set of released tuples T_r as follows.

Definition 4 (Exposure). *Let T_r be a set of released tuples over attributes A, let X and Y be two subsets of A such that $X \rightsquigarrow Y$, and let $y \in Y$ be a target. The* exposure *for y over T_r due to the dependency on X is $\mathcal{E}_r(X, y) = p_r(y)\Delta_r(X, y)$.*

Example 7. With reference to our running example, consider the evaluation of the exposure for target L_2, and suppose that $\Delta_r(\texttt{Age}, L_2) = 0.22$. If T_r is composed by 10 tuples on L_2 and 90 tuples of different locations, then $p_r(L_2) = 0.1$, and the exposure for L_2 is $\mathcal{E}_r(\texttt{Age}, L_2) = p_r(L_2)\Delta_r(\texttt{Age}, L_2) = 0.1 \cdot 0.22 = 0.02$. If, otherwise, T_r is composed by 10 L_2 tuples and 10 tuples of different locations, then $p_r(L_2) = 0.5$, and the exposure for L_2 is $\mathcal{E}_r(\texttt{Age}, L_2) = p_r(L_2)\Delta_r(\texttt{Age}, L_2) = 0.5 \cdot 0.22 = 0.11$.

Having characterized the exposure for y over a given release T_r, we now need to characterize when the release of a tuple t is safe or when the corresponding target $y = t[Y]$ is considered too exposed and the privacy of its associated sensitive information $s(y)$ at risk. Adapting Definition 4, we consider the release of a given target y safe if its exposure is not above the average exposure plus one standard deviation, $\sigma_{\mathcal{E}}$. The average exposure is $\frac{I_r(X,Y)}{|Y_r|}$, with $I_r(X, Y)$ the mutual information between attributes X and Y computed on T_r, and $|Y_r|$ the different values of Y in T_r. The average exposure is computed on T_r instead of T since the original table T is not known to external observers, who can only see and learn distributions from the released dataset T_r. Note that, clearly, the average exposure differs from the average of $\Delta(X, y)$ of Definition 3.

Definition 5 (Safe release). *Let T_r be a set of released tuples over attributes A, let X and Y be two subsets of A such that $X \rightsquigarrow Y$, let t be a tuple to be released, with $y = t[Y]$. The release of t is* safe *if $\mathcal{E}_{r'}(X, y) = p_{r'}(y) \cdot \Delta_{r'}(X, y)$ over $T_{r'} = T_r \cup t$ is less than $\frac{I_{r'}(X,Y)}{|Y_{r'}|} + \sigma_{\mathcal{E}}$, where $|\ Y_{r'}\ |$ is the number of different values of Y in $T_{r'}$.*

According to Definition 5, a tuple t, with $y = t[Y]$, can be released if the exposure $\mathcal{E}_{r'}(X, y)$ for $y = t[Y]$ over $T_{r'} = T_r \cup t$ (Definition 4) is less than the threshold $\frac{I_{r'}(X,Y)}{|Y_{r'}|} + \sigma_{\mathcal{E}}$.

5 Controlling exposure and regulating release

In the previous section we have characterized when a release is safe with respect to inference, which is when the distribution of values observable in the external world does not define the involved target as an X-outlier. The remaining aspect to consider is when to start enforcing such control. As a matter of fact, we are considering a scenario of incremental releases where the control needs to operate at run time and tuples can be requested one by one. We can clearly imagine that the release of the first few tuples will produce random distribution of values that will usually not resemble the actual distribution existing in the database, thus corresponding to an exposure of the different targets that can considerably differ from their real exposures. Typically, such a random exposure will characterize

the targets as X-outliers, thus blocking any release. Enforcing the control on the safe release at the start time of the system can therefore cause a denial of service in the system raising many false alarms (since also targets that are not X-outliers will have a random initial distribution that will differ from the baseline). In addition we note that clearly no observer could put confidence on statistics computed over a few releases as they cannot be considered accurate and their distribution can be completely random. There is therefore a starting time at which the data holder should allow the release of tuples regardless of whether the safety condition (Definition 5) is satisfied. After a sufficient amount of information has been released for a given target, subsequent releases should be controlled and allowed only if the release is safe. There is not a unique way to specify when the amount of information released should be considered sufficient. In the following, we discuss some possible approaches, which we are further investigating, performing experiments to evaluate their pros and cons in different settings.

- *Exposure accuracy.* A first approach consists in evaluating the accuracy of the exposure known to the observer with respect to the real exposure, which corresponds to the release of all the tuples of the target. Once the exposure approximates for the first time the real exposure (i.e., when the amount of released data is such that the corresponding exposure approximates the real exposure), the external knowledge can be considered accurate enough and a control can be triggered. Exposure accuracy is particular intuitive as a control on real X-outliers. In fact, exposure accuracy would trigger the control when the external observations would essentially leak the information that the target is close to its real distribution (which is a distribution corresponding to an X-outlier). Also, for targets which are not X-outliers it intuitively captures the fact that the external knowledge is not accurate.
- *Number of releases.* Another possible alternative solution is based on the number of tuples released for each given target. Intuitively this approach captures the fact that a limited number of tuples offers little knowledge to the observer since the distributions of values on them can be completely random and rarely correspond to the distribution actually existing in the database. The threshold on the number of tuples to be applied could be the same for the different targets or specific for each of them (e.g., targets with smaller occurrences could have a smaller threshold). The consideration of the number of released tuples naturally captures the confidence that the observer can put on the distributions based on the amount of data released: the more the data, the more the confident in the statistics.
- *Number of releases for different values of X.* While starting the control after a given number of tuples has been released for a given target can perform usually well, especially for targets that are not X-outliers, in few cases (and in particular for outliers) it may not suffice. For instance, with reference to our example, the first few tuples could all be referred to the same range value for `Age`, then exposing a peak for that range. To illustrate, consider our running example in Figure 2 and Figure 3. For our outlier location L_2, the

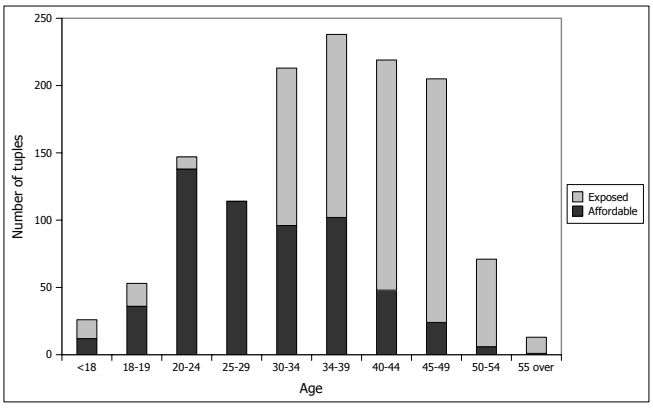


Fig. 4: Fitting the baseline distribution within the L_2-conditioned distribution

release of tuples in such a way that the distribution resembles the baseline distribution forces a maximum number of tuples that could be released for each age range. For instance, in the baseline distribution almost 19.67% of the soldiers are in the range [25-29], while in L_2 only 8.78% of tuples (140 tuples) fall in such range. Respecting the baseline distribution requires, even in the case where all tuples in the range [25-29] of L_2 are released to not release tuples in other ranges (so that the 140 tuples above actually correspond to 19.67%). Figure 4 graphically depicts this reasoning fitting the baseline distribution (in black) within the $L2$-conditioned distribution (gray going over the black). For each value range, no more than the number reached by the baseline distribution should be released.

The different approaches above have all an intuitive nature, providing different kinds of controls that perform well in different scenarios. They could therefore be applied individually or in conjunction to control releases of data in different settings. We have conducted some experiments to assess the impact and guarantees of the different types of controls. We have considered a table T as described in Example 1, where the 10000 tuples in the table have been randomly generated to respect the baseline distribution illustrated in Figure 3(a), which corresponds to the age distribution of the UK Regular Forces as at 1 March 2006. The distribution of the age ranges for each location are illustrated in Figure 2, characterizing location L_2 as the only `Age`-outlier. We performed five simulations, where each simulation consists in randomly releasing all the tuples in T. Before each simulation, the content of table T has been shuffled, so to produce different orders of release (and therefore different incremental observations over time). Figures 5(a)-(e) show for the five simulations how the exposure $\mathcal{E}_r(\texttt{Age}, L_i)$ varies with the number of released tuples for the five locations $L_1, \ldots, L_5$. The horizontal dashed lines represent the actual exposure $\mathcal{E}(\texttt{Age}, L_i)$ and the continuous lines represent the final value of the threshold (i.e., the threshold computed in correspondence of the last tuple released) introduced in Definition 5, which

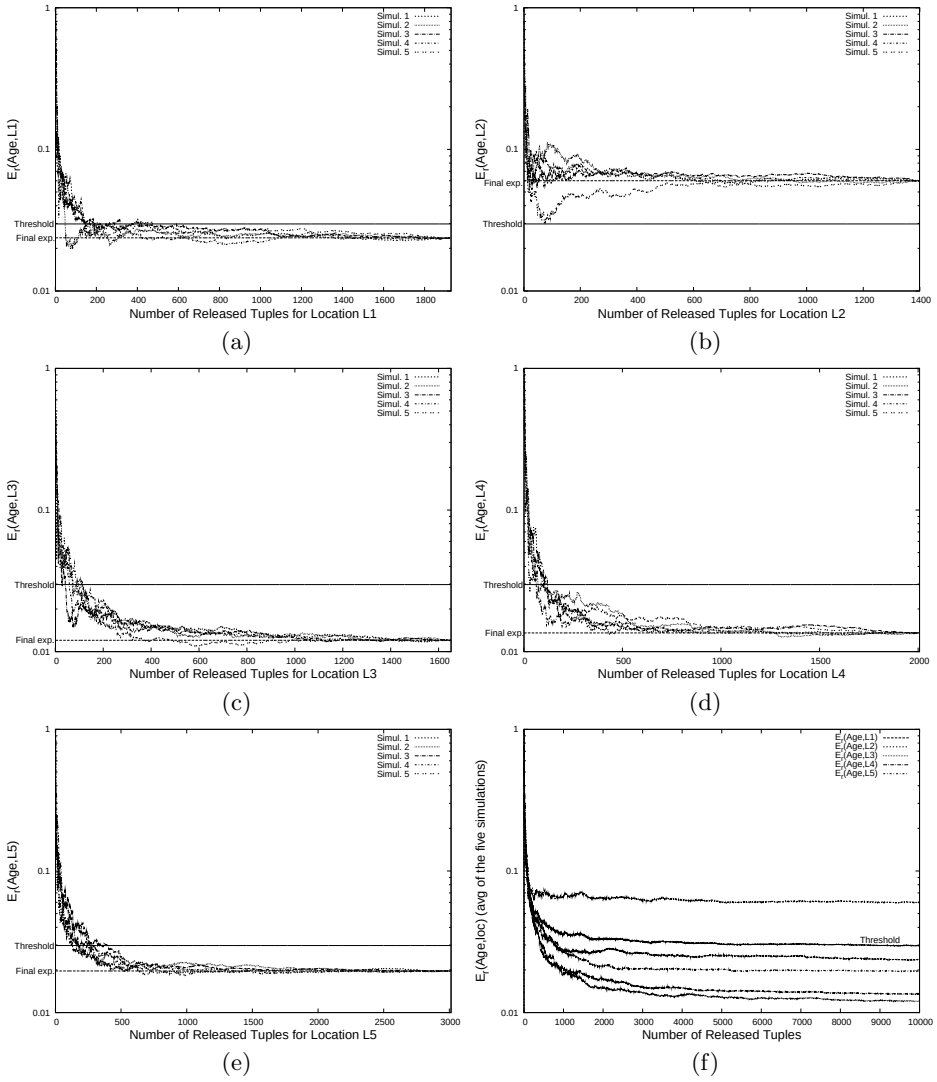


Fig. 5: Exposure variation for each location and simulation (a)-(e), and average on the five simulations of the exposure and threshold for each location (f)

varies as a new tuple is released. We report the final threshold value since in our experiments, after the release of a relatively small number of tuples (200 tuples on average), the released global distribution resembles the genuine distribution over table T, thus producing a threshold that quickly converges to the final value. Figure 5(f) illustrates, for each location, the average of the exposures $\mathcal{E}_r(\texttt{Age}, L_i)$ and of the threshold evaluated in the five simulations. In the graphs, we use a logarithmic scale for the ordinate axis to make the plot easier to view.

All the graphs in Figure 5 show that, for all locations and simulations, in the first releases there is a high exposure with fluctuations. The exposure then decreases and becomes stable (apart again small fluctuations) as the number of tuples released increases until the actual exposure is reached. The graphs confirm the intuition that the false positives happen mainly at the beginning of the releases (since distributions over a few tuples cannot be considered reliable). In fact, after a certain number of releases, for all locations but L_2 the exposure $\mathcal{E}_r(\texttt{Age}, L_i)$ computed by an observer typically is less than the threshold represented by the continuous horizontal lines, meaning that the tuples related to such locations can be safely released.

The different thresholds discussed above, when individually applied, have different impact on the control. In particular, the accuracy of the exposure with respect to actual exposure would start the control the first time the lines of the exposure of the releases come close to the actual exposure. While performing usually well and being intuitive, such a threshold has the side effect of not triggering the control for releases that show an anomalous distribution for targets that in fact are not X-outlier, that is, for false positives that remain such for a considerable number of releases. This is, for example, the case of the fifth simulated release for location L_1, where the exposure remains for a very long time above the threshold but the release is allowed since the exposure does not correctly reflect the actual exposure (in other words, it is a false positive). Whether such situation is legitimate or not depends on the kind of controls one wants to apply and whether releases of false positives should be considered as harmful. In such case, another threshold should be applied in alternative or in conjunction with the accuracy metrics. The threshold based on the number of tuples released would start the control after a given number of tuples are released, blocking any release considered unsafe according to our definition. In such case, in the specific case of the fifth simulation of location L_4, the release of tuples would be blocked after the threshold number of tuples has been reached.

6 Related work

Several research efforts have been recently dedicated to the problem of protecting privacy in data publication (e.g., [4,9,17,23]). In particular, considerable attention has been devoted to the problem of protecting the respondents' identities and the sensitive information associated with the respondents to whom the published data refer. Such proposals use the notion of k-anonymity [23] as a starting point or adopt some extensions of k-anonymity (e.g., [9,16,17,19]), and others are based on the idea of fragmenting data and publishing associations at the group level (e.g., [7,28]). Among them, t-closeness [17] and *(α_i, β_i)-closeness* [9] present some similarities with our work. t-closeness protects attribute disclosure by imposing that the distribution of sensitive values in the equivalence classes of the released table (i.e., in the groups of tuples with the same value for the quasi-identifying attributes) must be similar to the distribution in the private table. To this purpose, the t-closeness approach applies the Earth Mover's Distance

(EMD) for measuring the distance between the global distribution computed on the private table and the distributions computed within each equivalence classes. The distance between these distributions should be no more than t. In [9], the authors present an extension of t-closeness that overcomes some of its limitations (e.g., the difficulty in choosing a correct value for t and the impossibility to specify that there are some attribute values more sensitive than others). With this approach, the data publisher defines a different range $[\alpha_i, \beta_i]$ associated with each value v_i of a sensitive attribute. A released table is then acceptable when for each equivalence class the proportion of tuples in the class with a given sensitive value v_i falls in the corresponding range $[\alpha_i, \beta_i]$. Although our proposal and these two approaches have in common the fact that they consider inference issues caused by anomalous value distributions, our work addresses a different and more complex scenario characterized by incremental releases of detailed data. Also, in our scenario the sensitive information is not released but can be inferred due to a value distribution dependency between a set of attributes appearing in the released data and the sensitive property itself.

Inference problems have been studied extensively in the context of multilevel database systems (e.g., [15,18,20]). Most inference research addresses detection of inference channels within a database or at query processing time. In the first case, inference channels are removed by upgrading selected schema components or redesigning the schema (e.g., [22]). In the second case, database transactions are evaluated to determine whether they lead to illegal inferences and, if so, deny the query (e.g., [12,14,21,25]). Neither approach is however applicable to the problem under consideration. As a matter of fact, the inference problem we address is due to a dependency existing between the value distribution observable aggregating all the released tuples and the sensitive information that we want to protect. Previous work on inference focuses instead on locating inference channels based on semantic relationships between attributes or on queries submitted to the systems.

Our problem also has common aspects with the aggregation problem that arises when the aggregation of two or more data items is considered more sensitive than the single data items. A well-known example is the Secret Government Agency (SGA) Phonebook [24]: the entire phonebook is classified as confidential and it is accessible only by users with the appropriate clearance but single entries are unclassified and available to any requester. Although our problem is conceptually similar, the classical solutions developed for addressing the aggregation problem (e.g., [13,15,27]) are not directly applicable in our context. These approaches define a threshold on the amount of data that can be released to each user and focus on maintaining history and establishing how to control collusion among users.

Other related proposals are those used to assess the *interestingness* of association rules in knowledge discovery problems. In [26], the authors introduced the J-measure to assess the relevance of an association rule. In some sense, these proposals are complementary to ours, as they can be used for assessing dependencies among the attributes characterizing a data collection. The information

they produce can then be used as input to our approach for the definition of appropriate dependencies.

7 Conclusions

We considered the problem of protecting sensitive information in an incremental data release scenario, where the data holder releases non sensitive data on demand. As more and more data are released, an external observer can aggregate such data and infer the sensitive information by exploiting a dependency between the distribution of the non sensitive released data and the sensitive information itself. In this paper, we presented an approach for characterizing when data can be released without incurring to such inference. To this purpose, we defined when a distribution can be considered unusual and exploited for inference, and introduced the concept of safe release. Our work represents only a first step in the investigation of the problem and leaves space for further investigations, including: the experimental evaluations of the different approaches outlined in this paper for enforcing information release at run-time, the extension of the model to the consideration of inferences arising from information other than value distributions differing from a given pre-defined one, and the consideration of different types of knowledge that observers can exploit for inference.

Acknowledgments This work was supported in part by the EU within the 7FP project "PrimeLife" under grant agreement 216483 and by the Italian Ministry of Research within the PRIN 2008 project "PEPPER"(2008SY2PH4).

References

1. N.R. Adam and J.C. Wortmann. Security-control methods for statistical databases: A comparative study. *ACM Computing Surveys*, 21(4):515–556, December 1989.
2. C. Aggarwal and P.S. Yu, editors. *Privacy-Preserving Data Mining: Models and Algorithms*. Springer, 2008.
3. S. Cimato, M. Gamassi, V. Piuri, R. Sassi, and F. Scotti. Privacy-aware biometrics: Design and implementation of a multimodal verification system. In *Proc. of ACSAC 2008*, Anaheim, USA, Dec 2008.
4. V. Ciriani, S. De Capitani di Vimercati, S. Foresti, and P. Samarati. k-Anonymity. In T. Yu and S. Jajodia, editors, *Secure Data Management in Decentralized Systems*. Springer-Verlag, 2007.
5. V. Ciriani, S. De Capitani di Vimercati, S. Foresti, and P. Samarati. Microdata protection. In T. Yu and S. Jajodia, editors, *Secure Data Management in Decentralized Systems*. Springer-Verlag, 2007.
6. S. Dawson, S. De Capitani di Vimercati, P. Lincoln, and P. Samarati. Maximizing sharing of protected information. *Journal of Computer and System Sciences*, 64(3):496–541, May 2002.
7. S. De Capitani di Vimercati, S. Foresti, S. Jajodia, S. Paraboschi, and P. Samarati. Fragments and loose associations: Respecting privacy in data publishing. *Proc. of the VLDB Endowment*, 3(1), 2010.

8. R. M. Fano. *Transmission of Information; A Statistical Theory of Communications*. MIT University Press, New York, NY, USA, 1961.
9. K.B Frikken and Y. Zhang. Yet another privacy metric for publishing micro-data. In *Proc. of WPES 2008*, Alexandria, Virginia, USA, October 2008.
10. M. Gamassi, M. Lazzaroni, M. Misino, V. Piuri, D. Sana, and F. Scotti. Accuracy and performance of biometric systems. In *Proc. of IMTC 2004*, Como, Italy, 2004.
11. M. Gamassi, V. Piuri, D. Sana, and F. Scotti. Robust fingerprint detection for access control. In *Proc. of RoboCare Workshop 2005*, Rome, Italy, May 2005.
12. J.A. Goguen and J. Meseguer. Unwinding and inference control. In *Proc. of the IEEE Symp. on Security and Privacy*, Oakland, CA, USA, May 1984.
13. J.T. Haigh, R.C. O'Brien, and D.J. Thomsen. The LDV secure relational DBMS model. In S. Jajodia and C.E. Landwehr, editors, *Database Security, IV: Status and Prospects*, pages 265–279, North-Holland, 1991. Elsevier Science Publishers.
14. T.H. Hinke, H.S. Delugach, and A. Chandrasekhar. A fast algorithm for detecting second paths in database inference analysis. *Journal of Computer Security*, 3(2/3):147–168, 1995.
15. S. Jajodia and C. Meadows. Inference problems in multilevel secure database management systems. In *Information Security: an integrated collection of essays*, pages 570–584. IEEE Computer Sociery Press, 1995.
16. K. LeFevre, D.J. DeWitt., and R. Ramakrishnan. Mondrian multidimensional k-anonymity. In *Proc. of ICDE 2006*, Atlanta, GA, April 2006.
17. N. Li, Li T, and S. Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and ℓ-diversity. In *Proc. of ICDE 2007*, Istanbul, Turkey, April 2007.
18. T.F. Lunt. Aggregation and inference: facts and fallacies. In *Proc. of the IEEE Symp. on Security and Privacy*, Oakland, CA, USA, May 1989.
19. A. Machanavajjhala, J. Gehrke, and D. Kifer. ℓ-Diversity: Privacy beyond k-anonymity. In *Proc. of ICDE 2006*, Atlanta, GA, USA, April 2006.
20. D.G. Marks, A. Motro, and S. Jajodia. Enhancing the controlled disclosure of sensitive information. In *Proc. of ESORICS 1996*, Rome, Italy, September 1996.
21. M. Morgenstern. Controlling logical inference in multilevel database systems. In *Proc. of the IEEE Symp. on Security and Privacy*, Oakland, CA, USA, May 1988.
22. X. Qian, M.E. Stickel, P.D. Karp, T.F. Lunt, and T.D. Garvey. Detection and elimination of inference channels in multilevel relational database. In *Proc. of the IEEE Symp. on Research in Security and Privacy*, Oakland, CA, May 1993.
23. P. Samarati. Protecting respondents' identities in microdata release. *IEEE Transactions on Knowledge and Data Engineering*, 13(6):1010–1027, November/December 2001.
24. M. Schaefer (ed.). Multilevel data management security. Air Force Studies Board Committee on Multilevel Data Management Security, 1983.
25. G.W. Smith. Modeling security-relevant data semantics. *IEEE Transactions on Software Engineering*, 17(11):1195–1203, 1991.
26. P. Smyth and R.M. Goodman. An information theoretic approach to rule induction from databases. *IEEE Transactions on Knowledge and Data Engineering*, 4(4):301–316, August 1992.
27. D.D. Denning T.F. Lunt, R.R Schell, M. Heckman, and W.R. Shockley. The seaview security model. *IEEE Transactions of Software Engineering*, 16(6):593–607, June 1990.
28. X. Xiao and Y. Tao. Anatomy: Simple and effective privacy preservation. In *Proc. of VLDB 2006*, Korea, September 2006.